FUNDAÇÃO
GETULIO VARGAS

EPGE
Escola de Pós-Graduação
em Economia

Ensaios Econômicos

Escola de
Pós-Graduação
em Economia
da Fundação
Getulio Vargas

N° 713 ISSN 0104-8910

Model selection, estimation and forecasting in VAR models with short-run and long-run restrictions

, , ,

Janeiro de 2011

URL: http://hdl.handle.net/10438/7813

,
Model selection, estimation and forecasting in VAR models with short-run and long-run restrictions/ , , , - Rio de Janeiro : FGV,EPGE, 2011
?? p. - (Ensaios Econômicos; 713)

Inclui bibliografia.

CDD-330

Model selection, estimation and forecasting in VAR models with short-run and long-run restrictions

Athanasopoulos, George Guillén, Osmani Teixeira de Carvalho
Issler, João Victor Vahid, Farshid

Janeiro de 2011

Model selection, estimation and forecasting in VAR models with short-run and long-run restrictions

George Athanasopoulos
Department of Econometrics and Business Statistics
Monash University, Australia

Osmani Teixeira de Carvalho Guillén
Banco Central do Brasil and Ibmec-RJ
Rio de Janeiro, Brazil

João Victor Issler*
Graduate School of Economics – EPGE
Getulio Vargas Foundation
Rio de Janeiro, Brazil

Farshid Vahid
Department of Econometrics and Business Statistics
Monash University, Australia

Abstract

We study the joint determination of the lag length, the dimension of the cointegrating space and the rank of the matrix of short-run parameters of a vector autoregressive (VAR) model using model selection criteria. We suggest a new two-step model selection procedure which is a hybrid of traditional criteria and criteria with data-dependant penalties and we prove its consistency. A Monte Carlo study explores the finite sample performance of this procedure and evaluates the forecasting accuracy of models selected by this procedure. Two empirical applications confirm the usefulness of the model selection procedure proposed here for forecasting.

Keywords: Reduced rank models, model selection criteria, forecasting accuracy.

JEL Classification: C32, C53.

*Corresponding author: João Victor Issler, Graduate School of Economics – EPGE, Getulio Vargas Foundation, Praia de Botafogo 190 s. 1111, Rio de Janeiro, RJ 22253-900, Brazil. Tel: +5521 3799 5833, Fax: +5521 2553 8821, E-mail: Joao.Issler@fgv.br

1 Introduction

There is a large body of literature on the effect of cointegration on forecasting. Engle and Yoo (1987) compare the forecasts generated from an estimated vector error correction model (VECM) assuming that the lag order and the cointegrating rank are known, with those from an estimated VAR in levels with the correct lag. They find out that the VECM only produces forecasts with smaller mean squared forecast errors (MSFE) in the long-run. Clements and Hendry (1995) note that Engle and Yoo's conclusion is not robust if the object of interest is differences rather than levels, and use this observation to motivate their alternative measures for comparing multivariate forecasts. Hoffman and Rasche (1996) confirm Clements and Hendry's observation using a real data set. Christoffersen and Diebold (1998) also use Engle and Yoo's setup, but argue against using a VAR in levels as a benchmark on the grounds that the VAR in levels not only does not impose cointegration, it does not impose any unit roots either. Instead, they compare the forecasts of a correctly specified VECM with forecasts from correctly specified univariate models, and find no advantage in MSFE for the VECM. They use this result as a motivation to suggest an alternative way of evaluating forecasts of a cointegrated system. Silverstovs et al. (2004) extend Christoffersen and Diebold's results to multicointegrated systems. Since the afore-mentioned papers condition on the correct specification of the lag length and cointegrating rank, they cannot provide an answer as to whether we should examine the cointegrating rank of a system in multivariate forecasting if we do not have any a priori reason to assume a certain form of cointegration.

Lin and Tsay (1996) examine the effect on forecasting of the mis-specification of the cointegrating rank. They determine the lag order using the AIC, and compare the forecasting performance of estimated models under all possible numbers of cointegrating vectors (0 to 4) in a four-variable system. They observe that, keeping the lag order constant, the model with the correct number of cointegrating vectors achieves a lower MSFE for long-run forecasts, especially relative to a model that over-specifies the cointegrating rank. Although Lin and Tsay do not assume the correct specification of the lag length, their study also does not address the uncertainty surrounding the number of cointegrating vectors in a way that can lead to a modelling strategy for forecasting possibly cointegrated variables. Indeed, the results of their example with real data, in which they determine the cointegrating rank using a sequence of hypothesis tests, do not accord with their simulation results.

At the same time, there is an increasing amount of evidence of the advantage of considering rank restrictions for short-term forecasting in stationary VAR (and VARMA) models (see, for example, Ahn and Reinsel, 1988; Vahid and Issler, 2002; Athanasopoulos and Vahid, 2008). One feature of these papers is that they do not treat lag-length and rank uncertainty, differently. Their quest is to identify

the dimension of the most parsimonious state vector that can represent the dynamics of a system. Here, we add the cointegrating rank to the menu of unknowns and evaluate model selection criteria that determine all of these unknowns simultaneously. Our goal is to determine a modelling strategy that is useful for multivariate forecasting.

There are other papers in the literature that evaluate the performance of model selection criteria for determining lag-length and cointegrating rank, but they do not evaluate the forecast performance of the resulting models. Gonzalo and Pitarakis (1999) show that in large systems the usual model selection procedures may severely underestimate the cointegrating rank. Chao and Phillips (1999) show that the posterior information criterion (PIC) performs well in choosing the lag-length and the cointegrating rank simultaneously.

In this paper we evaluate the performance of model selection criteria in the simultaneous choice of the lag-length p, the rank of the cointegrating space q, and the rank of other parameter matrices r in a vector error correction model. We suggest a hybrid model selection strategy that selects p and r using a traditional model selection criterion, and then chooses q based on PIC. We then evaluate the forecasting performance of models selected using these criteria.

Our simulations cover the three issues of model building, estimation, and forecasting. We examine the performances of model selection criteria that choose p, r and q simultaneously (IC(p, r, q)), and compare their performances with a procedure that chooses p using a standard model selection criterion (IC(p)) and determines the cointegrating rank using a sequence of likelihood ratio tests proposed by Johansen (1988). We provide a comparison of the forecasting accuracy of fitted VARs when only cointegration restrictions are imposed, when cointegration and short-run restrictions are jointly imposed, and when neither are imposed. These comparisons take into account the possibility of model misspecification in choosing the lag length of the VAR, the number of cointegrating vectors, and the rank of other parameter matrices. In order to estimate the parameters of a model with both long-run and short-run restrictions, we propose a simple iterative procedure similar to the one proposed by Centoni et al. (2007).

It is very difficult to claim that any result found in a Monte Carlo study is general, especially in multivariate time series. There are examples in the VAR literature of Monte Carlo designs which led to all model selection criteria overestimating the true lag in small samples, therefore leading to the conclusion that the Schwarz criterion is the most accurate. The most important feature of these designs is that they have a strong propagation mechanism.[1] There are other designs with weak propagation mechanisms that result in all selection criteria underestimating the true lag and leading to the

[1] Our measure of the strength of the propagation mechanism is proportional to the trace of the product of the variance of first differences and the inverse of the variance of innovations.

conclusion that AIC's asymptotic bias in overestimating the true lag may actually be useful in finite samples (see Vahid and Issler, 2002, for references). We pay particular attention to the design of the Monte Carlo to make sure that we cover a wide range of data generating processes in terms of the strength of their propagation mechanisms.

The outline of the paper is as follows. In Section 2 we study finite VARs with long-run and short-run restrictions and motivate their empirical relevance. In Section 3, we outline an iterative procedure for computing the maximum likelihood estimates of parameters of a VECM with short-run restrictions. We provide an overview of model selection criteria in Section 4, and in particular we discuss model selection criteria with data dependent penalty functions. Section 5 describes our Monte Carlo design. Section 6 presents the simulation results and Section 8 concludes.

2 VAR models with long-run and short-run common factors

We start from the triangular representation of a cointegrated system used extensively in the cointegration literature (some early examples are Phillips and Hansen, 1990; Phillips and Loretan, 1991; Saikkonen, 1992). We assume that the K-dimensional time series

$$y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}, \quad t = 1, ..., T$$

where y_{1t} is $q \times 1$ (implying that y_{2t} is $(K-q) \times 1$) is generated from:

$$\begin{aligned} y_{1t} &= \beta y_{2t} + u_{1t} \\ \Delta y_{2t} &= u_{2t} \end{aligned} \tag{1}$$

where β is a $q \times (K-q)$ matrix of parameters, and

$$u_t = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

is a strictly stationary process with mean zero and positive definite covariance matrix. This is a data generating process (DGP) of a system of K cointegrated I(1) variables with q cointegrating vectors, also referred to as a system of K I(1) variables with $K-q$ common stochastic trends (some researchers also refer to this as a system of K variables with $K-q$ unit roots, which can be ambiguous if used out of context, and we therefore do not use it here).[2] The extra feature that we add to this fairly general DGP is that u_t is generated from a VAR of finite order p and rank r ($< K$).

[2]While in theory every linear system of K cointegrated I(1) variables with q cointegrating vectors can be represented in this way, in practice the decision on how to partition K-variables into y_{1t} and y_{2t} is not trivial, because y_{1t} are variables which must definitely have a non-zero coefficient in the cointegrating relationships.

In empirical applications, the finite VAR(p) assumption is routine. This is in contrast to the theoretical literature on testing for cointegration, in which u_t is assumed to be an infinite VAR, and a finite VAR(p) is used as an approximation (e.g. Saikkonen, 1992). Here, our emphasis is on building multivariate forecasting models rather than hypothesis testing. The finite VAR assumption is also routine when the objective is studying the maximum likelihood estimator of the cointegrating vectors, as in Johansen (1988).

The reduced rank assumption is considered for the following reasons. Firstly, this assumption means that all serial dependence in the K-dimensional vector time series u_t can be characterised by only $r < K$ serially dependent indices. This is a feature of most macroeconomic models, in which the short-run dynamics of the variables around their steady states are generated by a small number of serially correlated demand or supply shifters. Secondly, this assumption implies that there are $K - r$ linear combinations of u_t that are white noise. Gourieroux and Peaucelle (1992) call such time series "codependent," and interpret the white noise combinations as equilibrium combinations among stationary variables. This is justified on the grounds that, although each variable has some persistence, the white noise combinations have no persistence at all. For instance, if an optimal control problem implies that the policy instrument should react to the current values of the target variables, then it is likely that there will be such a linear relationship between the observed variables up to a measurement noise. Finally, many papers in multivariate time series literature provide evidence of the usefulness of reduced rank VARs for forecasting (see, for example, Velu et al., 1986; Ahn and Reinsel, 1988). Recently, Vahid and Issler (2002) have shown that failing to allow for the possibility of reduced rank structure can lead to developing seriously misspecified vector autoregressive models that produce bad forecasts.

The dynamic equation for u_t is therefore given by (all intercepts are suppressed to simplify the notation)

$$u_t = B_1 u_{t-1} + B_2 u_{t-2} + \cdots + B_p u_{t-p} + \varepsilon_t \tag{2}$$

where $B_1, B_2, ..., B_p$ are $K \times K$ matrices with $rank \left[\begin{array}{cccc} B_1 & B_2 & ... & B_p \end{array} \right] = r$, and ε_t is an i.i.d. sequence with mean zero and positive definite variance-covariance matrix and finite fourth moments. Note that the rank condition implies that each B_i has rank at most r, and the intersection of the null-spaces of all B_i is a subspace of dimension $K - r$. The following lemma derives the vector error correction representation of this data generating process.

Lemma 1 *The data generating process given by equations (1) and (2) has a reduced rank vector error correction representation of the type*

$$\Delta y_t = \gamma \left(\begin{array}{cc} I_q & -\beta \end{array} \right) y_{t-1} + \Gamma_1 \Delta y_{t-1} + \Gamma_2 \Delta y_{t-2} + \cdots + \Gamma_p \Delta y_{t-p} + \eta_t, \tag{3}$$

in which rank $\left[\begin{array}{cccc} \Gamma_1 & \Gamma_2 & ... & \Gamma_p \end{array}\right] \leq r$.

Proof. Refer to the working paper version of the current paper. ■

This lemma shows that the triangular DGP (1) under the assumption that the dynamics of its stationary component (i.e. u_t) can be characterised by a small number of common factors, is equivalent to a VECM in which the coefficient matrices of lagged differences have reduced rank and their left null-spaces overlap. Hecq et al. (2006) call such a structure a VECM with weak serial correlation common features (WSCCF).

We should note that the triangular structure (1) implies $K - q$ common Beveridge-Nelson (BN) trends, but the reduced rank structure assumed for u_t does not imply that deviations from the BN trends (usually refereed to as BN cycles) can be characterised as linear combinations of r common factors. Vahid and Engle (1993) show that a DGP with common BN trends and cycles is a special case of the above under some additional restrictions and therefore a stricter form of comevement. Hecq et al. (2006) show that the uncertainty in determining the rank of the cointegrating space can adversely affect inference on common cycles, and they conclude that testing for weak common serial correlation features is a more accurate means of uncovering short-run restrictions in vector error correction models.

Our objective is to come up with a model development methodology that allows for cointegration and weak serial correlation common features. For stationary time series, Vahid and Issler (2002) show that allowing for reduced rank models is beneficial for forecasting. For partially non-stationary time series, there is an added dimension of cointegration. Here, we examine the joint benefits of cointegration and short-run rank restrictions for forecasting partially non-stationary time series.

3 Estimation of VARs with short-run and long-run restrictions

The maximum likelihood estimation of the parameters of a VAR written in error-correction form

$$\Delta\, y_t = \Pi\; y_{t-1} + \Gamma_1 \Delta\, y_{t-1} + \Gamma_2 \Delta\, y_{t-2} + \cdots + \Gamma_p \Delta\, y_{t-p} + \eta_t \tag{4}$$

under the long-run restriction that the rank of Π is q, the short-run restriction that rank of $\left[\begin{array}{cccc} \Gamma_1 & \Gamma_2 & ... & \Gamma_p \end{array}\right]$ is r and the assumption of normality, is possible via a simple iterative procedure that uses the general principle of the estimation of reduced rank regression models (Anderson, 1951). Noting that the above model can be written as

$$\Delta\, y_t \;\; = \;\; \gamma\; \alpha' y_{t-1} + C\left[D_1 \Delta\, y_{t-1} + D_2 \Delta\, y_{t-2} + \cdots + D_p \Delta\, y_{t-p}\right] + \eta_t, \tag{5}$$

where α is a $K \times q$ matrix of rank q and C is a $K \times r$ matrix of rank r, one realises that if α was known, C and $D_i, i = 1, \ldots, p$, could be estimated using a reduced rank regression of $\Delta\, y_t$ on $\Delta\, y_{t-1}, \cdots, \Delta\, y_{t-p}$

after partialling out $\alpha' y_{t-1}$. Also, if $D_i, i = 1, \ldots, p$, were known, then γ and α could be estimated using a reduced rank regression of $\Delta\, y_t$ on y_{t-1} after controlling for $\sum_{i=1}^{p} D_i \Delta\, y_{t-i}$. This points to an easy iterative procedure for computing maximum likelihood estimates for all parameters.

Step 0. Estimate $[\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_p]$ from a reduced rank regression of $\Delta\, y_t$ on $(\Delta y_{t-1}, ..., \Delta y_{t-p})$ controlling for y_{t-1}. Recall that these estimates are simply coefficients of the canonical variates corresponding to the r largest squared partial canonical correlations (PCCs) between $\Delta\, y_t$ and $(\Delta y_{t-1}, ..., \Delta y_{t-p})$, controlling for y_{t-1}.

Step 1. Compute the PCCs between $\Delta\, y_t$ and y_{t-1} conditional on $[\hat{D}_1 \Delta\, y_{t-1} + \hat{D}_2 \Delta\, y_{t-2} + \cdots + \hat{D}_p \Delta\, y_{t-p}]$. Take the q canonical variates $\hat{\alpha}' y_{t-1}$ corresponding to the q largest squared PCCs as estimates of cointegrating relationships. Regress $\Delta\, y_t$ on $\hat{\alpha}' y_{t-1}$ and $[\hat{D}_1 \Delta\, y_{t-1} + \hat{D}_2 \Delta\, y_{t-2} + \cdots + \hat{D}_p \Delta\, y_{t-p}]$, and compute $\ln|\hat{\Omega}|$, the logarithm of the determinant of the residual variance matrix.

Step 2. Compute the PCCs between $\Delta\, y_t$ and $(\Delta y_{t-1}, ..., \Delta y_{t-p})$ conditional on $\hat{\alpha}' y_{t-1}$. Take the r canonical variates $[\hat{D}_1 \Delta\, y_{t-1} + \hat{D}_2 \Delta\, y_{t-2} + \cdots + \hat{D}_p \Delta\, y_{t-p}]$ corresponding to the largest r PCCs as estimates of $[D_1 \Delta\, y_{t-1} + D_2 \Delta\, y_{t-2} + \cdots + D_p \Delta\, y_{t-p}]$. Regress $\Delta\, y_t$ on $\hat{\alpha}' y_{t-1}$ and $[\hat{D}_1 \Delta\, y_{t-1} + \hat{D}_2 \Delta\, y_{t-2} + \cdots + \hat{D}_p \Delta\, y_{t-p}]$, and compute $\ln|\hat{\Omega}|$, the logarithm of the determinant of the residual variance matrix. If this is different from the corresponding value computed in Step 1, go back to Step 1. Otherwise, stop.

The value of $\ln|\hat{\Omega}|$ becomes smaller at each stage until it achieves its minimum, which we denote by $\ln|\hat{\Omega}_{p,r,q}|$. The values of $\hat{\alpha}$ and $[\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_p]$ in the final stage will be the maximum likelihood estimators of α and $[D_1, D_2, \ldots, D_p]$. The maximum likelihood estimates of other parameters are simply the coefficient estimates of the final regression. Note that although γ and α, and also C and $[D_1, D_2, \ldots, D_p]$, are only identified up to appropriate normalisations, the maximum likelihood estimates of Π and $[\Gamma_1, \Gamma_2, \ldots, \Gamma_p]$ are invariant to the choice of normalisation. Therefore, the normalisation of the canonical correlation analysis is absolutely innocuous, and the "raw" estimates produced from this procedure can be linearly combined to produce any desired alternative normalisation. Also, the set of variables that are partialled out at each stage should include constants and other deterministic terms if needed.

4 Model selection

The modal strategy in applied work for modelling a vector of I(1) variables is to use a model selection criterion for choosing the lag length of the VAR, then test for cointegration conditional on the lag-order,

and finally estimate the VECM. Hardly ever any further step is taken to simplify the model, and if any test of the adequacy of the model is undertaken, it is usually a system test. For example, to test the adequacy of the dynamic specification, additional lags of all variables are added to all equations, and a test of joint significance for K^2 parameters is used. For stationary time series, Vahid and Issler (2002) show that model selection criteria severely underestimate the lag order in weak systems, i.e. in systems where the propagation mechanism is weak. They also show that using model selection criteria (suggested in Lütkepohl, 1993, p. 202) to choose lag order and rank simultaneously can remedy this shortcoming significantly.

In the context of VECMs, one can consider selecting (p, r) with these model selection criteria first, and then use a sequence of likelihood ratio tests to determine the rank of the cointegrating space q. Specifically, these are the analogues of Akaike information criterion (AIC), the Hannan and Quinn criterion (HQ) and the Schwarz criterion (SC), and are defined as

$$AIC(p,r) = T\sum_{i=K-r+1}^{K}\ln\left(1-\hat{\lambda}_i^2(p)\right)+2(r(K-r)+rKp) \tag{6}$$

$$HQ(p,r) = T\sum_{i=K-r+1}^{K}\ln\left(1-\hat{\lambda}_i^2(p)\right)+2(r(K-r)+rKp)\ln\ln T \tag{7}$$

$$SC(p,r) = T\sum_{i=K-r+1}^{K}\ln\left(1-\hat{\lambda}_i^2(p)\right)+(r(K-r)+rKp)\ln T, \tag{8}$$

where K is the dimension of (number of series in) the system, r is the rank of $[\ \Gamma_1 \quad \Gamma_2 \quad ... \quad \Gamma_p\ ]$, p is the number of lagged differences in the VECM, T is the number of observations, and $\hat{\lambda}_i^2(p)$ are the sample squared PCCs between Δy_t and the set of regressors $(\Delta y_{t-1}, ..., \Delta y_{t-p})$ after the linear influence of y_{t-1} (and deterministic terms such as a constant term and seasonal dummies if needed) is taken away from them, sorted from the smallest to the largest. Traditional model selection criteria are special cases of the above when rank is assumed to be full, i.e. when r is equal to K. Here, the question of the rank of Π, the coefficient of y_{t-1} in the VECM, is set aside, and taking the linear influence of y_{t-1} away from the dependent variable and the lagged dependent variables concentrates the likelihood on $[\ \Gamma_1 \quad \Gamma_2 \quad ... \quad \Gamma_p\ ]$. Then, conditional on the p and the r that minimise one of these criteria, one can use a sequence of likelihood ratio tests to determine q. While in the proof of Theorem 2 we show that the estimators of p and r based on HQ and SC are consistent, the estimator of q from the sequential testing method with a fixed level of significance is obviously not. Moreover, the asymptotic distribution of the likelihood ratio test statistic for q conditional on selected p and r may be far from that when the true p and r are known (Leeb and Potscher, 2005). Here, we study model selection criteria which choose p, r and q.

We consider two classes of model selection criteria. First, we consider direct extensions of the AIC, HQ and SC to the case where the rank of the cointegrating space, which is the same as the rank of Π, is also a parameter to be selected by the criteria. Specifically, we consider

$$AIC(p,r,q) = T\ln|\hat{\Omega}_{p,r,q}| + 2(q(K-q)+Kq+r(K-r)+rKp) \quad (9)$$

$$HQ(p,r,q) = T\ln|\hat{\Omega}_{p,r,q}| + 2(q(K-q)+Kq+r(K-r)+rKp)\ln\ln T \quad (10)$$

$$SC(p,r,q) = T\ln|\hat{\Omega}_{p,r,q}| + (q(K-q)+Kq+r(K-r)+rKp)\ln T, \quad (11)$$

where $\ln|\hat{\Omega}_{p,r,q}|$ (the minimised value of the logarithm of the determinant of the variance of the residuals of the VECM of order p, with Π having rank q and $[\ \Gamma_1 \quad \Gamma_2 \quad ... \quad \Gamma_p\ ]$ having rank r) is computed by the iterative algorithm described above in Section 3. Obviously, when $q = 0$ or $q = K$, we are back in the straightforward reduced rank regression framework, where one set of eigenvalue calculations for each p provides the value of the log-likelihood function for $r = 1, ..., K$. Similarly, when $r = K$, we are back in the usual VECM estimation, and no iterations are needed.

We also consider a model selection criterion with a data dependent penalty function. Such model selection criteria date back at least to Poskitt (1987), Rissanen (1987) and Wallace and Freeman (1987). The model selection criterion that we consider in this paper is closer to those inspired by the "minimum description length (MDL)" criterion of Rissanen (1987) and the "minimum message length (MML)" criterion of Wallace and Freeman (1987). Both of these criteria measure the complexity of a model by the minimum length of the uniquely decipherable code that can describe the data using the model. Rissanen (1987) establishes that the closest the length of the code of any emprical model can possibly get to the length of the code of the true DGP P_θ is at least as large as $\frac{1}{2}\ln|E_\theta(\mathrm{FIM}_M(\hat{\theta}))|$, where $\mathrm{FIM}_M(\hat{\theta})$ is the Fisher information matrix of model M (i.e., $[-\partial^2 lnl_M/\partial\theta\partial\theta']$, the second derivative of the log-likelihood function of the model M) evaluated at $\hat{\theta}$, and E_θ is the mathematical expectation under P_θ. Rissanen uses this bound as a penalty term to formulate the MDL as a model selection criterion,

$$\mathrm{MDL} = -\ln l_M(\hat{\theta}) + \frac{1}{2}\ln|\mathrm{FIM}_M(\hat{\theta})|.$$

Wallace and Freeman's MML is also based on coding and information theory but is derived from a Bayesian perspective. The MML criterion is basically the same as the MDL plus an additional term that is the prior density of the parameters evaluated at $\hat{\theta}$ (see Wallace, 2005, for more details and a summary of recent advances in this line of research). While the influence of this term is dominated by the other two terms as sample size increases, it plays the important role of making the criterion invariant to arbitrary linear transformations of the regressors in a regression context.

Based on their study of the asymptotic form of the Bayesian data density, Phillips (1996) and Phillips and Ploberger (1996) design the posterior information criterion (PIC), which is similar to

MML and MDL criteria. Their important contribution has been to show that such criteria can be applied to partially nonstationary time series as well.[3] Chao and Phillips (1999) use the PIC for simultaneous selection of the lag length and cointegration rank in VARs.

There are practical difficulties in working with PIC that motivates simplifying this criterion. One difficulty is that $\text{FIM}_M(\hat{\theta})$ must be derived and coded for all models considered (The details of the Fisher information matrix for a reduced rank VECM is given in the appendix). A more important one is the large dimension of $\text{FIM}_M(\hat{\theta})$. For example, if we want to choose the best VECM allowing for up to 4 lags in a six variable system, we have to compute determinants of square matrices of dimensions as large as 180. These calculations are likely to push the boundaries of numerical accuracy of computers, in particular when these matrices are ill-conditioned[4]. This, and the favourable results of the HQ criterion in selecting lag p and rank of stationary dynamics r, led us to consider a two step procedure.

4.1 A two-step procedure for model selection

In the first step, the linear influence of y_{t-1} is removed from Δy_t and $(\Delta y_{t-1}, ..., \Delta y_{t-p})$, then $\text{HQ}(p, r)$, as defined in (7), is used to determine p and r. Then PIC is calculated for the chosen values of p and r, for all q from 0 to K. This reduces the task to $K + 1$ determinant calculations only.

Theorem 2 *If the data generating process is*

$$\Delta y_t = c + \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \Gamma_2 \Delta y_{t-2} + \cdots + \Gamma_{p_0} \Delta y_{t-p_0} + \eta_t$$

in which

(i) all roots of the characteristic polynomial of the implied VAR for y_t are on or outside the unit circle and all those on the unit circle are $+1$;

(ii) the rank of Π is $q_0 \leq K$, which implies that Π can be written as $\gamma\alpha'$ where γ and α are full rank $K \times q_0$ matrices;

(iii) $\gamma'_\perp \left(I - \sum_{i=1}^{p_0} \Gamma_i\right) \alpha_\perp$ has full rank where $\gamma_\perp$ and $\alpha_\perp$ are full rank $K \times (K - q_0)$ matrices such that $\gamma'_\perp \gamma = \alpha'_\perp \alpha = 0$;

(iv) the rank of $[\ \Gamma_1 \quad \Gamma_2 \quad ... \quad \Gamma_{p_0}\ ]$ *is $r_0 \leq K$;*

(v) the rank of Γ_{p_0} is not zero;

(vi) $E(\eta_t \mid \mathcal{F}_{t-1}) = 0$ and $E(\eta_t \eta'_t \mid \mathcal{F}_{t-1}) = \Omega$ positive definite where $\mathcal{F}_{t-1}$ is the σ-field generated by

[3] Ploberger and Phillips (2003) generalised Rissanen's result to show that even for trending time series, the distance between any empirical model and the P_θ is larger or equal to $\frac{1}{2} \ln |E_\theta(\text{FIM}_M)|$ almost everywhere on the parameter space. They use the outer-product formulation of the information matrix, which has the same expected value as the negative of the second derivative under P_θ.

[4] In our simulations, we came across one case where the determinant was returned to be a small negative number even though the matrix was symmetric positive definite. This happened both using GAUSS and also using MATLAB.

$\{\eta_{t-1}, \eta_{t-2}, \ldots\}$, *and* $E(\eta_{it}^4) < \infty$ *for* $i = 1, 2, \ldots, K$,
and the maximum possible lag considered $p_{\max} \geq p_0$, *then the estimators of* p, r *and* q *obtained from the two step procedure explained above are consistent.*

Proof. See Appendix B. ■

5 Monte-Carlo design

To make the Monte-Carlo simulation manageable, we use a three-dimensional VAR. We consider VARs in levels with lag lengths of 2 and 3, which translates to 1 and 2 lagged differences in the VECM. This choice allows us to study the consequences of both under- and over-parameterisation of the estimated VAR.

For each p_0, r_0 and q_0 we draw many sets of parameter values from the parameter space of cointegrated VARs with serial correlation common features that generate difference stationary data. In order to ensure that the DGPs considered do not lie in a subset of the parameter space that implies only very weak or only very strong propagation mechanisms we choose 50 DGPs with system R^2s (as defined in Vahid and Issler, 2002) that range between 0.3 and 0.65, with a median between 0.4 and 0.5 and 50 DGPs with system R^2s that range between 0.65 and 0.9, with a median between 0.7 and 0.8.

From each DGP, we generate 1,000 samples of 100, 200 and 400 observations (the actual generated samples were longer, but the initial part of each generated sample is discarded to reduce the effect of initial conditions). In summary, our results are based on 1,000 samples of 100 different DGPs — a total of 100,000 different samples — for each of $T = 100$, 200 or 400 observations.

The Monte-Carlo procedure can be summarised as follows. Using each of the 100 DGPs, we generate 1,000 samples (with 100, 200 and 400 observations). We record the lag length chosen by traditional (full-rank) information criteria, labeled IC(p) for IC={AIC, HQ, SC}, and the corresponding lag length chosen by alternative information criteria, labeled IC(p, r, q) for IC={AIC, HQ, SC, PIC, HQ-PIC} where the last is the hybrid procedure we propose in Section 4.1.

We should note that although we present the results averaged over all 100 DGPs we have also analysed the results for the DGPs with low and high R^2s separately. We indeed found that any advantage of model selection criteria with a relatively smaller (larger) penalty factor was accentuated when only considering DGPs with relatively weaker (stronger) propagation mechanisms. In order to save space we do not present these results here but they are available upon request.

For choices made using the traditional IC(p) criteria, we use Johansen's (1988, 1991) trace test at the 5% level of significance to select q, and then estimate a VECM with no short-run restrictions. For choices made using IC(p, r, q), we use the two step procedure of Section 4.1 to obtain the triplet

(p, r, q), and then estimate the resulting VECM with SCCF restrictions using the algorithm of Section 3. For each case we record the out-of-sample forecasting accuracy measures for up to 16 periods ahead. We then compare the out-of-sample forecasting accuracy measures for these two types of VAR models.

5.1 Measuring forecast accuracy

We measure the accuracy of forecasts using the traditional trace of the mean-squared forecast error matrix (TMSFE) and the determinant of the mean-squared forecast error matrix |MSFE| at different horizons. We also compute Clements and Hendry's (1993) *generalized forecast error second moment* (GFESM). GFESM is the determinant of the expected value of the outer product of the vector of stacked forecast errors of all future times up to the horizon of interest. For example, if forecasts up to h quarters ahead are of interest, this measure will be:

$$\text{GFESM} = \left| E \begin{pmatrix} \tilde{\varepsilon}_{t+1} \\ \tilde{\varepsilon}_{t+2} \\ \vdots \\ \tilde{\varepsilon}_{t+h} \end{pmatrix} \begin{pmatrix} \tilde{\varepsilon}_{t+1} \\ \tilde{\varepsilon}_{t+2} \\ \vdots \\ \tilde{\varepsilon}_{t+h} \end{pmatrix}' \right|,$$

where $\tilde{\varepsilon}_{t+h}$ is the K-dimensional forecast error of our K-variable model at horizon h. This measure is invariant to elementary operations that involve different variables (TMSFE is not invariant to such transformations), and also to elementary operations that involve the same variable at different horizons (neither TMSFE nor |MSFE| is invariant to such transformations). In our Monte-Carlo, the above expectation is evaluated for every model, by averaging over replications.

There is one complication associated with simulating 100 different DGPs. Simple averaging across different DGPs is not appropriate, because the forecast errors of different DGPs do not have identical variance-covariance matrices. Lütkepohl (1985) normalises the forecast errors by their true variance-covariance matrix in each case before aggregating. Unfortunately, this would be a very time consuming procedure for a measure like GFESM, which involves stacked errors over many horizons. Instead, for each information criterion, we calculate the percentage gain in forecasting measures, comparing the full-rank models selected by IC(p), with the reduced-rank models chosen by IC(p, r, q). This procedure is done at every iteration and for every DGP, and the final results are then averaged.

6 Monte-Carlo simulation results

6.1 Selection of lag, rank, and the number of cointegrating vectors

Simulation results are reported in "three-dimensional" frequency tables. The columns correspond to the percentage of times the selected models had cointegrating rank smaller than the true rank ($q < q_0$), equal to the true rank ($q = q_0$) and larger than the true rank ($q > q_0$). The rows correspond to similar

information about the rank of short-run dynamics r. Information about the lag-length is provided within each cell, where the entry is disaggregated on the basis of p. The three numbers provided in each cell, from left to right, correspond to percentages with lag lengths smaller than the true lag, equal to the true lag and larger than the true lag. The 'Total' column on the right margin of each table provides information about marginal frequencies of p and r only. The row titled 'Total' on the bottom margin of each table provides information about the marginal frequencies of p and q only. Finally, the bottom right cell provides marginal information about the lag-length choice only.

We report results of two sets of 100 DGPs. Table 1 summarises the model selection results for 100 DGPs that have one lag in differences with a short-run rank of one and cointegrating rank of two, i.e., $(p_0, r_0, q_0) = (1, 1, 2)$. Table 2 summarises the model selection results for 100 DGPs that have two lags in differences with a short-run rank of one and cointegrating rank of one $(p_0, r_0, q_0) = (2, 1, 1)$. These two groups of DGPs are contrasting in the sense that the second group of DGPs have more severe restrictions in comparison to the first one.

The first three panels of the tables correspond to all model selection based on the traditional model selection criteria. The additional bottom row for each of these three panels provides information about the lag-length and the cointegrating rank, when the lag-length is chosen using the simple version of that model selection criterion and the cointegrating rank is chosen using the Johansen procedure, and in particular the sequential trace test with 5% critical values that are adjusted for sample size. Comparing the rows labeled 'AIC+J', 'HQ+J' and 'SC+J', we conclude that the inference about q is not sensitive to whether the selected lag is correct or not. In Table 1 all three criteria choose the correct q approximately 54%, 59% and 59% of the time for sample sizes 100, 200 and 400, respectively. In Table 2 all three criteria choose the correct q approximately 70%, 82% and 82% of the time for sample sizes 100, 200 and 400, respectively.

From the first three panels of Table 1 we can clearly see that traditional model selection criteria do not perform well in choosing p, r and q jointly in finite samples. The percentages of times the correct model is chosen are only 22%, 26% and 29% with the AIC, 39%, 52% and 62% with HQ, and 42%, 63% and 79% with SC, for sample sizes of 100, 200 and 400, respectively. Note that when we compare the marginal frequencies of (p, r), HQ is the most successful for choosing both p and r, a conclusion that is consistent with results in Vahid and Issler (2002).

The main reason for not being able to determine the triplet (p, r, q) correctly is the failure of these criteria to choose the correct q. Ploberger and Phillips (2003) show that the correct penalty for free parameters in the long-run parameter matrix is larger than the penalty considered by traditional model selection criteria. Accordingly, all three criteria are likely to over-estimate q in finite samples, and of them SC is likely to appear relatively most successful because it assigns a larger penalty to all free

parameters, even though the penalty is still less than ideal. This is exactly what the simulations reveal.

The fourth panel of Table 1 includes results for the PIC. The percentages of times the correct model is chosen increase to 52%, 77% and 92% for sample sizes of 100, 200 and 400, respectively. Comparing the margins, it becomes clear that this increased success relative to HQ and SC is almost entirely due to improved precision in the selection of q. The PIC chooses q correctly 76%, 91% and 97% of the time for sample sizes 100, 200 and 400, respectively. Furthermore, for the selection of p and r only, PIC does not improve upon HQ.

Similar conclusions can be reached from the results for the $(2, 1, 1)$ DGPs presented in Table 2. We note that in this case, even though the PIC improves on HQ and SC in choosing the number of cointegrating vectors, it does not improve on HQ or SC in choosing the exact model, because it severely underestimates p. This echoes the findings of Vahid and Issler (2002) in the stationary case that the Schwarz criterion (recall that the PIC penalty is of the same order as the Schwarz penalty in the stationary case) severely underestimates the lag length in small samples in reduced rank VARs.

Our Monte-Carlo results show that the advantage of PIC over HQ and SC is in the determination of the cointegrating rank. Indeed, HQ seems to have an advantage over PIC in selecting the correct p and r in small samples. These results coupled with the practical difficulties in computing the PIC we outline in Section 4 motivated us to consider the two-step alternative procedure to improve the model selection task.

The final panels in Tables 1 and 2 summarise the performance of our two-step procedure. In both tables we can see that the hybrid HQ-PIC procedure improves on all other criteria in selecting the exact model. The improvement is a consequence of the advantage of HQ in selecting p and r better, and PIC in selecting q better.

Note that our hybrid procedure results in over-parameterised models more often than just using PIC as the model selection criterion. We examined whether this trade-off has any significant consequences for forecasting and found that it does not. In all simulation settings, models selected by the hybrid procedure with HQ-PIC as the model selection criteria forecast better than models selected by PIC. Again, we do not present these results here, but they are also available upon request.

6.2 Forecasts

Recall that the forecasting results are expressed as the percentage improvement in forecast accuracy measures of possibly rank reduced models over the unrestricted VAR model in levels selected by SC. Also, note that the object of interest in this forecasting exercise is assumed to be the first difference of variables, although GFESM gives a measure of accuracy that is the same for levels or differences.

We label the models chosen by the hybrid procedure proposed in the previous section and estimated

by the iterative process of Section 3 as VECM(HQ-PIC). We label the models estimated by the usual Johansen method with AIC as the model section criterion for the lag order as VECM(AIC+J).

Table 3 presents the forecast accuracy improvements in a $(1, 1, 2)$ setting. In terms of the trace and determinant of the MSFE matrix, there is some improvement in forecasts over unrestricted VAR models at all horizons. With only 100 observations, GFESM worsens for horizons 8 and longer. This means that if the object of interest was some combination of differences across different horizons (for example, the levels of all variables or the levels of some variables and first differences of others), there may not have been any improvement in the MSFE matrix. With 200 or more observations, all forecast accuracy measures show some improvement, with the more substantial improvements being for the one-step-ahead forecasts. Also note that the forecasts of the models selected by the hybrid procedure are almost always better than those produced by the model chosen by the AIC plus Johansen method, which only pays attention to lag-order and long-run restrictions.

Table 4 presents the forecast accuracy improvements in a $(2, 1, 1)$ setting. This set of DGPs have more severe rank reductions than the $(1, 1, 2)$ DGPs, and, as a result, the models selected by the hybrid procedure show more substantial improvements in forecasting accuracy over the VAR in levels, in particular for smaller sample sizes. Forecasts produced by the hybrid procedure are also substantially better than forecasts produced by the AIC+Johansen method, which does not incorporate short-run rank restrictions. Note that although the AIC+Johansen forecasts are not as good as the HQ-PIC forecasts, they are substantially better than the forecasts from unrestricted VARs at short horizons.

Following a request from a referee in Tables 3 and 4 we have also presented Diebold and Mariano (1995) tests for equal predictive accuracy between the rank reduced specifications and the unrestricted VARs for the TMSFE. In general the results are as expected. Models that incorporate reduced rank restrictions rarely forecast significantly worse than the unrestricted models. They either perform the same or significantly better than the unrestricted VARs.

7 Empirical example

The techniques discussed in this paper are applied in two different forecasting exercises to two data sets. The first data set contains Brazilian inflation, as measured by three different types of consumer-price indices, available on a monthly basis from 1994:9 to 2009:11, with a span of more than 15 years (183 observations). It was extracted from IPEADATA – a public database with downloadable Brazilian data (http://www.ipeadata.gov.br/). The second data set consists of real U.S. per-capita private output[5], personal consumption per-capita, and fixed investment per-capita, available on a quarterly basis from

[5]Private output is GNP minus federal government's consumption and investment.

1947:1 to 2009:3, with a span of more than 62 years (251 observations). It was extracted from FRED's database of the Federal Reserve Bank of St. Louis (http://research.stlouisfed.org/fred2/). Considering that we will keep some observations for forecast evaluation (90 observations), the size of these data bases are close to the number of simulated observations in the Monte-Carlo exercise for $T = 100$ and $T = 200$ respectively.

7.1 Forecasting Brazilian Inflation

The Brazilian data set consists of three alternative measures of consumer price indices. The first is the official consumer price index used in the Brazilian Inflation-Targeting Program. It is computed by IBGE, the statistics bureau of the Brazilian government, labeled here as CPI-IBGE. The second is the consumer price index computed by Getulio Vargas Foundation, a traditional private institution which computes several Brazilian price indices since 1947, labeled here as CPI-FGV. The third is the consumer price index computed by FIPE, an institute of the Department of Economics of the University of São Paulo, labeled here as CPI-FIPE.

These three indices capture different aspects of Brazilian consumer-price inflation. First, they differ in terms of geographical coverage. CPI-FGV is based on prices in 12 different metropolitan areas in Brazil, 11 of which are also covered by CPI-IBGE[6]. On the other hand, CPI-FIPE only covers São Paulo – the largest city in Brazil – also covered by the other two indices. Tracked consumption bundles are also different across indices. CPI-FGV is based on the typical consumption bundles of consumers with income between 1 and 33 minimum wages. CPI-IBGE covers consumption baskets of consumers with income between 1 and 40 minimum wages, while CPI-FIPE focuses on consumers with income between 1 and 20 minimum wages.

Although all three indices measure consumer-price inflation in Brazil, Granger-causality tests confirm the usefulness of conditioning on alternative indices to forecast any given index in the models estimated here. We compare the forecasting performance of (i) the VAR in log-levels, with lag length chosen by the standard Schwarz criterion; (ii) the VECM, using standard AIC for choosing the lag length and Johansen's test for choosing the cointegrating rank; and (iii) the reduced rank model, with rank and lag length chosen simultaneously using the Hannan-Quinn criterion and cointegrating rank chosen using PIC. All forecast comparisons are made using the first difference of the logarithms of the price indices, i.e., price inflation.

For all three models, the estimation sample starts from 1994:9 through 2001:2, with 78 observations. With these initial estimates, we compute the applicable choices of p, r, and q for each model and forecast inflation up to 16 months ahead. Keeping the initial observation fixed (1994:9), we add one

[6]There are no metropolitan areas covered by CPI-IBGE that are not covered by CPI-FGV.

observation at the end of the estimation sample, choose potentially different values for p, r, and q for each model, and forecast inflation again up to 16 months ahead. This procedure is then repeated until the final estimation sample reaches 1994:9 through 2008:7, with 167 observations. Then, we have a total of 90 out-of-sample forecasts for each horizon (1 to 16 months ahead), which are used for forecast evaluation. Thus, the estimation sample varies from 78 to 167 observations and mimics closely the simulations labeled $T = 100$ in the Monte-Carlo exercise.

Results of the exercise described above are presented in Table 5. For all horizons, there are substantial forecasting gains of the VECM(HQ-PIC) over the VAR in levels: for example, for 12 months (one year) ahead, TMSFE, |MSFE| and GFESM show gains of 33.6%, 38.4% and 120.3% respectively. The VECM(AIC+J) forecasts are also better than the VAR in levels forecasts, but the improvements are not as large. The comparison between VECM(HQ-PIC) and VECM(AIC+J) shows gains for the former everywhere.

Table 5 also includes the results of Diebold-Mariano tests for equality of mean squared errors of each pair of forecasts for each individual series for the reported horizons. These are reported using three comma separated symbols (one for each series) in parentheses below the TMSFE values. Each symbol indicates if the null hypothesis of the equality of mean squared forecast errors is rejected in favor of a one sided alternative and if so the level of significance at which it is rejected. The results indicate that in this application, the VECM(HQ-PIC) forecast of inflation based on every one of the three series has significantly lower mean squared error than the corresponding forecast from the VAR in log-levels. The test for the equality of the mean squared forecast errors of the VECM(HQ-PIC) and the VECM(AIC+J) rejects equality in favor of better VECM(HQ-PIC) forecasts at horizons 1, 4 and 8. It should be noted that there is no case where either the VAR in log-levels or the VECM(AIC+J) generate a significantly smaller MSE vis-à-vis the VECM(HQ-PIC) for any of the inflation series at any horizon.

It is also worth reporting the choices of p, r, and q for the best models studied here as the estimation sample increases from 1994:9-2001:2 all the way to 1994:9-2008:7. While the VECM(HQ-PIC) chose $p = 1$, $r = 1$ or 2, and $q = 0$, most of the time (on the rare occasion it chose $p = 3$, $q = 1$), the VECM(AIC+J) chose $p = 1$, $q = 1$, most of the time (on rare occasions it chose $p = 5$, $q = 0$ or $q = 3$). Hence, the superior performance of the VECM(HQ-PIC) vis-à-vis the VECM(AIC+J) may be due to either imposing a reduced-rank structure or to ignoring potential cointegration relationships. This is especially true for the shorter horizons. If the coverage of the price indices were similar, then one would expect a single common trend (i.e. two cointegrating vectors) in this system. However, the definition and coverage of these indices are substantially different and our analysis shows that this creates very persistent differences in these series and suggests that the users of these series must pay

careful attention to their definitions and choose the appropriate one for their purpose. Even if one believes that these persistent differences appear to be non-mean-reverting because of the short span of the data and they will eventually die out, our analysis shows that for forecasting purposes these differences are persistent enough that is better to model them as unit roots rather than stationary processes[7]. This is consistent with the results of Stock (1996).

7.2 Forecasting U.S. Macroeconomic Aggregates

The data set consists of logarithms of real U.S. per-capita private output – y, personal consumption per-capita – c, and fixed investment per-capita – i, extracted from FRED's database on a quarterly frequency[8] from 1947:1 to 2009:3.

Again, we compare the forecasting performance of (i) the VAR in log-levels, with lag length chosen by the standard Schwarz criterion; (ii) the VECM, using standard AIC for choosing the lag length and Johansen's test for choosing the cointegrating rank; and (iii) the reduced rank model, with rank and lag length chosen simultaneously using the Hannan-Quinn criterion and cointegrating rank chosen using PIC, estimated by the iterative process of Section 3. All forecast comparisons are made using the first difference of the log-levels of the data, i.e., using $\Delta \log(y_t)$, $\Delta \log(c_t)$, and $\Delta \log(i_t)$. For all three models, the first estimation sample covers the period 1947:1 to 1983:2, a total of 146 observations. As before, we keep expanding the estimation sample until it reaches 1947:1 to 2005:3, with 235 observations. This produces a total of 90 out-of-sample forecasts for each horizon that are used for forecast evaluation. Since the estimation sample varies from 146 to 235 observations, it corresponds closely to the simulations labeled $T = 200$ in the Monte-Carlo exercise.

Results of the exercise described above are presented in Table 6. For all horizons, there are considerable forecasting gains for the VECM(HQ-PIC) over the VAR in levels: at 4 quarters (one year) ahead, TMSFE, |MSFE| and GFESM show gains of 56.3%, 83.5% and 134.7% respectively. The forecasting gains of the VECM(AIC+J) over the VAR in levels, though statistically significant, are not as large especially for the short-run horizons. The comparison between VECM(HQ-PIC) and VECM(AIC+J) shows gains for the former in one quarter to four quarters ahead forecasts. The Diebold-Mariano tests for equality of the mean squared forecast errors for each of the three series provide evidence that HQ-PIC forecasts have significantly smaller mean squared errors than the VECM(AIC+J) forecasts

[7]We imposed $q = 2$ and repeated the forecasting exercise. The resulting forecasts were not even as good as VECM(AIC+J) forecasts. Detailed results are not reported here to save space, but are available to interested readers.

[8]Using FRED's mnemonics (2010) for the series, the precise definitions are: PCECC96 - consumption, FPIC96 - investment, and (GNP96 - FGCEC96) - output. Population series mnemonics is POP, which is only available from 1952 on in FRED. To get a complete series starting in 1947:1 it was spliced with the same series available in DRI database, whose mnemonics is GPOP.

for horizons 1 and 4.

Finally, we investigate the final choices of p, r, and q as the estimation sample increases from 1947:1-1983:2 to 1947:1-2005:3. For the VECM(HQ-PIC) they are: $p = 1$, $r = 2$, and $q = 0$, everywhere, while the VECM(AIC+J) chose $p = 1$ half of time and $p = 3$ the other half and $q = 0$ half of the time and $q = 1$ the other half. As in the previous example, the selected cointegrating ranks may not accord with theoretical priors. A theoretical real business cycle model hypothesizes that the technology shocks is the driver of the only common stochastic trend in all real variables and hence implies that y, c and i have two cointegrating vectors. What we learn from the data though is that even if this theory is correct, one or both of these cointegrating relationships must have such a high persistence (roots close to unity) that for forecasting purposes it is best if they are modeled as unit roots. If we impose $q = 2$ the forecasts are even inferior to VECM(AIC+J) forecasts (detailed results not reported to save space).

8 Conclusion

Motivated by the results of Vahid and Issler (2002) on the success of the Hannan-Quinn criterion in selecting the lag length and rank in stationary VARs, and the results of Ploberger and Phillips (2003) and Chao and Phillips (1999) on the generalisation of Rissanen's theorem to trending time series and the success of PIC in selecting the cointegrating rank in VARs, we propose a combined HQ-PIC procedure for the simultaneous choice of the lag-length and the ranks of the short-run and long-run parameter matrices in a VECM and we prove its consistency. Our simulations show that this procedure is capable of selecting the correct model more often than other alternatives such as pure PIC or SC.

In this paper we also present forecasting results that show that models selected using this hybrid procedure produce better forecasts than unrestricted VARs selected by SC and cointegrated VAR models whose lag length is chosen by the AIC and whose cointegrating rank is determined by the Johansen procedure. We have chosen these two alternatives for forecast comparisons because we believe that these are the model selection strategies that are most often used in the empirical literature. However, we have considered several other alternative model selection strategies and the results are qualitatively the same: the hybrid HQ-PIC procedure leads to models that generally forecast better than VAR models selected using other procedures.

A conclusion we would like to highlight is the importance of short-run restrictions for forecasting. We believe that there has been much emphasis in the literature on the effect of long-run cointegrating restrictions on forecasting. Given that long-run restrictions involve the rank of only one of the parameter matrices of a VECM, and that inference on this matrix is difficult because it involves inference about stochastic trends in variables, it is puzzling that the forecasting literature has paid so

much attention to cointegrating restrictions and relatively little attention to lag-order and short-run restrictions in a VECM. The present paper fills this gap and highlights the fact that the lag-order and the rank of short-run parameter matrices are also important for forecasting. Our hybrid model selection procedure and the accompanying simple iterative procedure for the estimation of a VECM with long-run and short-run restrictions provide a reliable methodology for developing multivariate autoregressive models that are useful for forecasting.

How often restrictions of the type considered in this paper are present in VAR approximations to real life data generating processes is an empirical question. Macroeconomic models in which trends and cycles in all variables are generated by a small number of dynamic factors fit in this category. Also, empirical papers that study either regions of the same country or similar countries in the same region often find these kinds of long-run and short-run restrictions. We illustrate the usefulness of the model-selection strategy discussed above in two empirical applications: forecasting Brazilian inflation and U.S. macroeconomic aggregates growth rates. We find gains of imposing short- and long-run restrictions in VAR models, since the VECM(HQ-PIC) and the VECM(AIC+J) outperform the VAR in levels everywhere. Tests of equal variance confirm that these gains are significant. Moreover, ignoring short-run restrictions usually produce inferior forecasts with these data, since the VECM(HQ-PIC) outperforms the VECM(AIC+J) almost everywhere, but these gains are not always significant in tests of equal variance.

It is true that discovering the "true" model is a different objective from model selection for forecasting. However, in the context of partially non-stationary variables, there are no theoretical results that lead us to a definite model selection strategy for forecasting. Using a two variable example, Elliott (2006) shows that, ignoring estimation uncertainty, whether or not considering cointegration will improve short-run or long-run forecasting depends on all parameters of the DGP, even the parameters of the covariance matrix of the errors. In addition there is no theory that tells us whether finite sample biases of parameter estimates will help or hinder forecasting in partially non-stationary VARs. Given this state of knowledge, when one is given the task of selecting a single model for forecasting it is reasonable to use a model selection criterion that is more likely to pick the "true" model and in this paper we verify that VARs selected by our hybrid model selection strategy are likely to produce better forecasts than unrestricted VARs and VARs that only incorporate cointegration restrictions.

Acknowledgements

We would like to thank the Associate Editor, two anonymous referees, Heather Anderson, Giovanni Forchini, Taya Dumrongrittikul, Yin Liao, Shuping Shi and Wenying Yao for useful comments and suggestions, and Claudia F. Rodrigues for excellent research assistance. João Issler thanks CNPq, FAPERJ and INCT for financial support. George Athanasopoulos and Farshid Vahid acknowledge support from the Australian Research Council grant DP0984399.

References

Ahn, S. K. and G. C. Reinsel (1988). Nested reduced-rank autoregressive models for multiple time series. *Journal of the American Statistical Association 83*, 849–856.

Anderson, H. M. and F. Vahid (1998). Testing multiple equation systems for common nonlinear components. *Journal of Econometrics 84*, 1–36.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics 22*, 327–351.

Athanasopoulos, G. and F. Vahid (2008). VARMA versus VAR for macroeconomic forecasting. *Journal of Business and Economic Statistics 26*, 237–252.

Aznar, A. and M. Salvador (2002). Selecting the rank of the cointegation space and the form of the intercept using an information criterion. *Econometric Theory 18*, 926–947.

Centoni, M., G. Cubbada, and A. Hecq (2007). Common shocks, common dynamics and the international business cycle. *Economic Modelling 24*, 149–166.

Chao, J. C. and P. C. B. Phillips (1999). Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics 91*, 227–271.

Christoffersen, P. F. and F. X. Diebold (1998). Cointegration and long-horizon forecasting. *Journal of Business and Economic Statistics 16*, 450–458.

Clements, M. P. and D. F. Hendry (1993). On the limitations of comparing mean squared forecast errors (with discussion). *Journal of Forecasting 12*, 617–637.

Clements, M. P. and D. F. Hendry (1995). Forecasting in cointegrated systems. *Journal of Applied Econometrics 10*, 127–146.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*, 253–263.

Elliott, G. (2006). Forecasting with trending data. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, Chapter 11, pp. 555–604. Elsevier.

Engle, R. F. and S. Yoo (1987). Forecasting and testing in cointegrated systems. *Journal of Econometrics 35*, 143–159.

Gonzalo, J. and J. Pitarakis (1995). Specification via model selection in vector error correction models. *Economic Letters 60*, 321–328.

Gonzalo, J. and J. Pitarakis (1999). Dimensionality effect in cointegration tests. In R. Engle and H. White (Eds.), *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, Chapter 9, pp. 212–229. New York: Oxford University Press.

Gourieroux, C. and I. Peaucelle (1992). Series codependantes application a l'hypothese de parite du pouvoir d'achat. *Revue d'Analyse Economique 68*, 283–304.

Hecq, A., F. Palm, and J.-P. Urbain (2006). Common cyclical features analysis in VAR models with cointegration. *Journal of Econometrics 132*, 117–141.

Hoffman, D. L. and R. H. Rasche (1996). Assessing forecast performance in a cointegrated system. *Journal of Applied Econometrics 11*, 495–517.

Johansen, S. (1988). Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control 12*, 231–254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica 59*, 1551–1580.

Leeb, H. and B. M. Potscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*, 21–59.

Lin, J. L. and R. S. Tsay (1996). Cointegration constraints and forecasting: An empirical examination. *Journal of Applied Econometrics 11*, 519–538.

Lütkepohl, H. (1985). Comparison of criteria for estimating the order of a vector autoregressive process. *Journal of Time Series Analysis 9*, 35–52.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis* (2nd ed.). Berlin-Heidelberg: Springer-Verlag.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: John Wiley and Sons.

Paulsen, J. (1984). Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis 5*, 115–127.

Phillips, P. C. B. (1996). Econometric model determination. *Econometrica 64*, 763–812.

Phillips, P. C. B. and B. Hansen (1990). Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies 57*, 99–125.

Phillips, P. C. B. and M. Loretan (1991). Estimating long-run economic equilibria. *Review of Economic Studies 58*, 407–436.

Phillips, P. C. B. and W. Ploberger (1996). An asymptotic theory of Bayesian inference for time series. *Econometrica 64*, 381–413.

Ploberger, W. and P. C. B. Phillips (2003). Empirical limits for time series econometric models. *Econometrica 71*, 627–673.

Poskitt, D. S. (1987). Precision, complexity and bayesian model determination. *Journal of the Royal Statistical Society B 49*, 199–208.

Quinn, B. G. (1980). Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society B 42*, 182–185.

Reinsel, G. C. (1997). *Elements of Multivariate Time Series* (2nd ed.). New York: Springer-Verlag.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society B 49*, 223–239.

Saikkonen, P. (1992). Estimation and testing of cointegrated systems by an autoregressive approximation. *Econometric Theory 8*, 1–27.

Silverstovs, B., T. Engsted, and N. Haldrup (2004). Long-run forecasting in multicointegrated systems. *Journal of Forecasting 23*, 315–335.

Sims, C. A., J. H. Stock, and M. W. Watson (1990). Inference in linear time series models with some unit roots. *Econometrica 58*, 113–144.

Stock, J. H. (1996). VAR, error correction and pretest forecasts at long horizons. *Oxford Bulletin of Economics and Statistics 58*, 685–701.

Tsay, R. S. (1984). Order selection in nonstationary autoregressive models. *Annals of Statistics 12*, 1425–1433.

Vahid, F. and R. F. Engle (1993). Common trends and common cycles. *Journal of Applied Econometrics 8*, 341–360.

Vahid, F. and J. V. Issler (2002). The importance of common cyclical features in VAR analysis: A Monte-Carlo study. *Journal of Econometrics 109*, 341–363.

Velu, R. P., G. C. Reinsel, and D. W. Wickern (1986). Reduced rank models for multiple time series. *Biometrika 73*, 105–118.

Wallace, C. (2005). *Statistical and Inductive Inference by Minimum Message Length*. Berlin: Springer.

Wallace, C. and P. Freeman (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society B 49*, 240–265.

A The Fisher information matrix of the reduced rank VECM

Assuming that the first observation in the sample is labeled observation $-p+1$ and that the sample contains $T+p$ observations, we write the K-variable reduced rank VECM

$$\Delta y_t = \gamma' \begin{pmatrix} I_q & \beta' \end{pmatrix} y_{t-1} + \begin{pmatrix} I_r \\ C' \end{pmatrix} [D_1 \Delta y_{t-1} + D_2 \Delta y_{t-2} + \cdots + D_p \Delta y_{t-p}] + \mu + e_t,$$

or in stacked form

$$\Delta Y = Y_{-1} \begin{pmatrix} I_q \\ \beta \end{pmatrix} \gamma + WD \begin{pmatrix} I_r & C \end{pmatrix} + \iota_T \mu' + E,$$

where

$$\begin{aligned}
\underset{T\times K}{\Delta Y} &= \begin{bmatrix} \Delta y_1' \\ \vdots \\ \Delta y_T' \end{bmatrix}, \quad \underset{T\times K}{Y_{-1}} = \begin{bmatrix} y_0' \\ \vdots \\ y_{T-1}' \end{bmatrix}, \quad \underset{T\times K}{E} = \begin{bmatrix} e_1' \\ \vdots \\ e_T' \end{bmatrix} \\
\underset{T\times Kp}{W} &= \begin{pmatrix} \Delta Y_{-1} & \cdots & \Delta Y_{-p} \end{pmatrix} = \begin{bmatrix} \Delta y_0' & \cdots & \Delta y_{-p+1}' \\ \vdots & \vdots & \vdots \\ \Delta y_{T-1}' & \cdots & \Delta y_{T-p}' \end{bmatrix} \\
\underset{Kp\times r}{D} &= \begin{pmatrix} D_1' \\ \vdots \\ D_p' \end{pmatrix},
\end{aligned}$$

and ι_T is a $T \times 1$ vector of ones. When e_t are $N(0, \Omega)$ and serially uncorrelated, the log-likelihood function, conditional on the first p observations being known, is:

$$\begin{aligned} \ln l(\theta, \omega) &= -\frac{KT}{2}\ln(2\pi) - \frac{T}{2}\ln|\Omega| - \frac{1}{2}\sum_{t=1}^{T} e_t'\Omega^{-1}e_t \\ &= -\frac{KT}{2}\ln(2\pi) - \frac{T}{2}\ln|\Omega| - \frac{1}{2}tr\left(E\Omega^{-1}E'\right), \end{aligned}$$

where

$$\theta = \begin{pmatrix} vec(\beta) \\ vec(\gamma) \\ vec(D) \\ vec(C) \\ \mu \end{pmatrix}$$

is a $(K-q)q + Kq + Kpr + r(K-r) + K$ matrix of mean parameters, and $\omega = vech(\Omega)$ is a $K(K+1)/2$ vector of unique elements of the variance matrix. The differential of the log-likelihood is (see Magnus and Neudecker, 1988)

$$\begin{aligned} d\ln l(\theta, \omega) &= -\frac{T}{2}tr\Omega^{-1}d\Omega + \frac{1}{2}tr\left(\Omega^{-1}d\Omega\Omega^{-1}E'E\right) - \frac{1}{2}tr\left(\Omega^{-1}E'dE\right) - \frac{1}{2}tr\left(\Omega^{-1}dE'E\right) \\ &= \frac{1}{2}tr\left(\Omega^{-1}\left(E'E - T\Omega\right)\Omega^{-1}d\Omega\right) - tr\left(\Omega^{-1}E'dE\right), \end{aligned}$$

and the second differential is:

$$\begin{aligned} d^2\ln l(\theta, \omega) &= tr\left(d\Omega^{-1}\left(E'E - T\Omega\right)\Omega^{-1}d\Omega\right) + \frac{1}{2}tr\left(\Omega^{-1}\left(2E'dE - Td\Omega\right)\Omega^{-1}d\Omega\right) \\ &\quad -tr\left(d\Omega^{-1}E'dE\right) - tr\left(\Omega^{-1}dE'dE\right). \end{aligned}$$

Since we eventually want to evaluate the Fisher information matrix at the maximum likelihood estimator, and at the maximum likelihood estimator $\hat{E}'\hat{E} - T\hat{\Omega} = 0$, and also $\hat{\Omega}^{-1}\hat{E}'dE/d\theta = 0$ (these are apparent from the first differentials), we can delete these terms from the second differential, and use $tr(AB) = vec(A')'vec(B)$ to obtain

$$\begin{aligned} d^2\ln l(\theta, \omega) &= -\frac{T}{2}tr\left(\Omega^{-1}d\Omega\Omega^{-1}d\Omega\right) - tr\left(\Omega^{-1}dE'dE\right) \\ &= -\frac{T}{2}(d\omega)'\mathbf{D}_K'\left(\Omega^{-1}\otimes\Omega^{-1}\right)\mathbf{D}_K d\omega - (vec(dE))'\left(\Omega^{-1}\otimes I_T\right)vec(dE), \end{aligned}$$

where $\mathbf{D}_K$ is the "duplication matrix". From the model, we can see that

$$dE = -Y_{-1}\begin{pmatrix} 0 \\ d\beta \end{pmatrix}\gamma - Y_{-1}\begin{pmatrix} I_q \\ \beta \end{pmatrix}d\gamma - WdD\begin{pmatrix} I_r & C \end{pmatrix} - WD\begin{pmatrix} 0 & dC \end{pmatrix} - \iota_T d\mu',$$

and therefore

$$vec(dE) = -\begin{bmatrix} \gamma'\otimes Y_{-1}^{(2)} & I_K\otimes Y_{-1}\begin{pmatrix} I_q \\ \beta \end{pmatrix} & \begin{pmatrix} I_r \\ C' \end{pmatrix}\otimes W & \begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix}\otimes WD & I_K\otimes\iota_T \end{bmatrix} d\theta.$$

Hence, the elements of the Fisher information matrix are:

$$
\begin{aligned}
FIM_{11} &= \gamma\Omega^{-1}\gamma' \otimes Y_{-1}^{(2)\prime}Y_{-1}^{(2)}, \quad FIM_{12} = \gamma\Omega^{-1} \otimes Y_{-1}^{(2)\prime}Y_{-1} \begin{pmatrix} I_q \\ \beta \end{pmatrix}, \\
FIM_{13} &= \gamma\Omega^{-1} \begin{pmatrix} I_r \\ C' \end{pmatrix} \otimes Y_{-1}^{(2)\prime}W, \quad FIM_{14} = \gamma\Omega^{-1} \begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix} \otimes Y_{-1}^{(2)\prime}WD \\
FIM_{15} &= \gamma\Omega^{-1} \otimes Y_{-1}^{(2)\prime}\iota_T \\
FIM_{22} &= \Omega^{-1} \otimes \begin{pmatrix} I_q & \beta' \end{pmatrix} Y_{-1}'Y_{-1} \begin{pmatrix} I_q \\ \beta \end{pmatrix}, \quad FIM_{23} = \Omega^{-1} \begin{pmatrix} I_r \\ C' \end{pmatrix} \otimes \begin{pmatrix} I_q & \beta' \end{pmatrix} Y_{-1}'W \\
FIM_{24} &= \Omega^{-1} \begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix} \otimes \begin{pmatrix} I_q & \beta' \end{pmatrix} Y_{-1}'WD, \quad FIM_{25} = \Omega^{-1} \otimes \begin{pmatrix} I_q & \beta' \end{pmatrix} Y_{-1}'\iota_T \\
FIM_{33} &= \begin{pmatrix} I_r & C \end{pmatrix} \Omega^{-1} \begin{pmatrix} I_r \\ C' \end{pmatrix} \otimes W'W, \quad FIM_{34} = \begin{pmatrix} I_r & C \end{pmatrix} \Omega^{-1} \begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix} \otimes W'WD \\
FIM_{35} &= \begin{pmatrix} I_r & C \end{pmatrix} \Omega^{-1} \otimes W'\iota_T \\
FIM_{44} &= \begin{pmatrix} 0 & I_{K-r} \end{pmatrix} \Omega^{-1} \begin{pmatrix} 0 \\ I_{K-r} \end{pmatrix} \otimes D'W'WD, \quad FIM_{45} = \begin{pmatrix} 0 & I_{K-r} \end{pmatrix} \Omega^{-1} \otimes D'W'\iota_T \\
FIM_{55} &= \Omega^{-1} \otimes \iota_T'\iota_T = \Omega^{-1} \times T
\end{aligned}
$$

B Proof of Theorem 2

The first three assumptions ensure that Δy_t is covariance stationary and y_t are cointegrated with cointegrating rank q_0. These together with assumption (vi) ensure that all sample means and covariances of Δy_t consistently estimate their population counterparts and the least squares estimator of parameters is consistent. Assumptions (iv) and (v) state that the true rank is r_0 and the true lag-length is p_0 (or the lag order of the implied VAR in levels is p_0+1). For any (p,r) pair, the second step of the analysis produces the least squares estimates of $\Gamma_1,\ldots,\Gamma_p$ with rank r when no restrictions are imposed on Π (Anderson, 1951). Reinsel (1997) contains many of the results that we use in this proof. Under the assumption of normality, these are the ML estimates of $\Gamma_1,\ldots,\Gamma_p$ with rank r with Π unrestricted and the resulting $\hat{\Omega}_{p,r}$ used in the HQ procedure is the corresponding ML estimate of Ω. Note that normality of the true errors is not needed for the proof. We use the results of Sims et al. (1990) who show that in the above model the least squares estimates of $\Gamma_1,\ldots,\Gamma_p$ have the standard asymptotic properties as in stationary VARs, in particular that they consistently estimate their population counterparts and that their rate of convergence is the same as $T^{-\frac{1}{2}}$. Let $z_t, z_{t-1},\ldots,z_{t-p}$ denote $\Delta y_t, \Delta y_{t-1}, ..., \Delta y_{t-p}$ after the influence of the constant and y_{t-1} is removed from them and let $Z, Z_{-1},\ldots,Z_{-p}$ denote $T\times K$ matrices with $z_t', z_{t-1}',\ldots,z_{t-p}'$ in their row $t=1,\ldots,T$ (we assume that the sample starts from $t=-p_{\max}+1$), and let $W_p=[Z_{-1}\vdots\cdots\vdots Z_{-p}]$ and $B_p=[\Gamma_1\vdots\cdots\vdots\Gamma_p]'$. The estimated model in the second step can be written as:

$$Z = W_p\hat{B}_p + \hat{U}_p$$

where $\hat{U}_p$ is the $T \times K$ matrix of residuals when the lag length is p. In an unrestricted regression $\ln|\frac{1}{T}\hat{U}_p'\hat{U}_p| = \ln|\frac{1}{T}(Z'Z - Z'W_p(W_p'W_p)^{-1}W_p'Z)| = \ln|\frac{1}{T}Z'Z| + \ln|I_K - (Z'Z)^{-1}Z'W_p(W_p'W_p)^{-1}W_p'Z|$ $= \ln|\frac{1}{T}Z'Z| + \sum_{i=1}^{K}\ln(1-\hat{\lambda}_i^2(p))$, where $\hat{\lambda}_1^2(p) \leq \hat{\lambda}_2^2(p) \leq \ldots \leq \hat{\lambda}_K^2(p)$, the eigenvalues of $(Z'Z)^{-1}Z'W_p(W_p'W_p)^{-1}W_p'Z$ are the ordered sample partial canonical correlations between Δy_t and $\Delta y_{t-1}, ..., \Delta y_{t-p}$ after the influence of a constant and y_{t-1} has been removed. Under the restriction that the rank of B is r, the log-determinant of the squared residuals matrix becomes $\ln|\frac{1}{T}\hat{U}_{p,r}'\hat{U}_{p,r}| = \ln|\frac{1}{T}Z'Z| + \sum_{i=K-r+1}^{K}\ln(1-\hat{\lambda}_i^2(p))$. Further, note that $W_p = [W_{p-1}\vdots Z_{-p}]$ and from the geometry of least squares we know

$Z'W_p(W_p'W_p)^{-1}W_p'Z = Z'W_{p-1}(W_{p-1}'W_{p-1})^{-1}W_{p-1}'Z + Z'Q_{p-1}Z_{-p}(Z_{-p}'Q_{p-1}Z_{-p})^{-1}Z_{-p}'Q_{p-1}Z$ where $Q_{p-1} = I_T - W_{p-1}(W_{p-1}'W_{p-1})^{-1}W_{p-1}'$.

(i) Consider $p = p_0$ and $r = r_0 - 1$: $\ln|\frac{1}{T}\hat{U}_{p_0,r_0-1}'\hat{U}_{p_0,r_0-1}| - \ln|\frac{1}{T}\hat{U}_{p_0,r_0}'\hat{U}_{p_0,r_0}| = -\ln(1-\hat{\lambda}_{K-r_0+1}^2(p_0))$. $\hat{\lambda}_{K-r_0+1}^2(p_0)$ converges in probability to its population counterpart, the r_0-th largest eigenvalue of $\Sigma_z^{-1}B_{p_0}'\Sigma_w B_{p_0}$, where Σ_x denotes the population second moment of the vector x. This population canonical correlation is strictly greater than zero because B_{p_0} has rank r_0. Therefore

$p\lim\,(\ln|\frac{1}{T}\hat{U}_{p_0,r_0-1}'\hat{U}_{p_0,r_0-1}| - \ln|\frac{1}{T}\hat{U}_{p_0,r_0}'\hat{U}_{p_0,r_0}|) = -\ln(1-\lambda_{K-r_0+1}^2(p_0)) > 0$.

(ii) Consider $p = p_0 - 1$ and $r = r_0$:

$$\begin{aligned}(Z'Z)^{-1}Z'W_{p_0}(W_{p_0}'W_{p_0})^{-1}W_{p_0}'Z &= (Z'Z)^{-1}Z'W_{p_0-1}(W_{p_0-1}'W_{p_0-1})^{-1}W_{p_0-1}'Z \\ &\quad +(Z'Z)^{-1}Z'Q_{p_0-1}Z_{-p_0}(Z_{-p_0}'Q_{p_0-1}Z_{-p_0})^{-1}Z_{-p_0}'Q_{p_0-1}Z.\end{aligned}$$

Since the second matrix on the right side is positive semi-definite, it follows that $\hat{\lambda}_i^2(p_0-1) \leq \hat{\lambda}_i^2(p_0)$ for all $i = 1, \ldots, K$.[9] We know that the probability limits of the smallest $K - r_0$ eigenvalues $\hat{\lambda}_i^2(p_0)$ are zero. Therefore, the probability limits of the smallest $K - r_0$ eigenvalues $\hat{\lambda}_i^2(p_0-1)$ must also be zero. Moreover, the trace of the matrix on the left is equal to the sum of the traces of the two matrices on the right of the equal sign. The probability limit of the last matrix on the right side is $\Sigma_z^{-1}\Gamma_{p_0}'\Sigma_{z.w}\Gamma_{p_0}$ where $\Sigma_{z.w} = p\lim(\frac{1}{T}Z_{-p_0}'Q_{p_0-1}Z_{-p_0})$, and since $rank(\Gamma_{p_0}) > 0$ by assumption, the probability limit of the trace of the second matrix on the right hand side will be strictly positive (note that even when $\Gamma_{p_0}$is nilpotent (i.e. has all zero eigenvalues even though its rank is not zero), $\Sigma_z^{-1}\Gamma_{p_0}'\Sigma_{z.w}\Gamma_{p_0}$ will not be nil-potent). Therefore it must be that $p\lim\hat{\lambda}_i^2(p_0-1) < p\lim\hat{\lambda}_i^2(p_0)$ for at least one $i = r_0 + 1, \ldots, K$. This implies that $p\lim\,(\ln|\frac{1}{T}\hat{U}_{p_0-1,r_0}'\hat{U}_{p_0-1,r_0}| - \ln|\frac{1}{T}\hat{U}_{p_0,r_0}'\hat{U}_{p_0,r_0}|) = \sum_{i=K-r_0+1}^{K}(\ln(1-\lambda_i^2(p_0-1)) - \ln(1-\lambda_i^2(p_0))) > 0$.

[9]Some textbooks define positive definiteness and associated inequalities concerning ordered eigenvalues for symmetric matrices only. Note that since the eigenvalues of any square matrix A is the same as the eigenvalues of GAG^{-1} for any invertible matrix G with the same dimensions as A (see Magnus and Neudecker, 1988, Chapter 1) one can choose $G = (Z'Z)^{\frac{1}{2}}$ and make all matrices on both sides of the inequality symmetric without changing any of their eigenvalues. Indeed this is a useful transformation for calculating canonical correlations because computer procedures for computation of eigenvalues of symmetric matrices are more accurate than those for general matrices.

(i) and (ii), together with the fact that $|\hat{U}'_{p_1,r_1}\hat{U}_{p_1,r_1}| \geq |\hat{U}'_{p_2,r_2}\hat{U}_{p_2,r_2}|$ whenever $p_1 \leq p_2$ and $r_1 \leq r_2$ (i.e., for all nested models the less restrictive cannot fit worse) imply that the probability limit of $\ln|\frac{1}{T}\hat{U}'_{p_0,r_0}\hat{U}_{p_0,r_0}|$ is *strictly smaller* than the probability limit of $\ln|\frac{1}{T}\hat{U}'_{p,r}\hat{U}_{p,r}|$ for all ($p \leq p_0$ and $r < r_0$) or ($p < p_0$ and $r \leq r_0$). Although the penalty favours the smaller models, the reward for parsimony increases at rate $\ln\ln T$ while the reward for better fit increases at rate T and therefore dominates. Hence, the probability of choosing a model with ($p \leq p_0$ and $r < r_0$) or ($p < p_0$ and $r \leq r_0$) goes to zero asymptotically.

(i') In (i), replace $p = p_0$ with $p = \tilde{p} \geq p_0$. The model now includes redundant lags whose true coefficients are zero and these coefficients are consistently estimated. Moreover, adding these zero parameters does not change the rank. Therefore all arguments in (i) apply to this case also, and we can therefore deduce that the probability of under-estimating r with this procedure goes to zero asymptotically.

(ii') In (ii), replace $r = r_0$ with $r = \tilde{r} \geq r_0$. The model now does not impose all rank restrictions that the true data generating process includes, but the extra eigenvalues will converge to their true value of zero asymptotically and all arguments in (ii) apply to this case also. Therefore, we can conclude that the probability of under-estimating p with this procedure goes to zero asymptotically.

(iii) Consider $p = \tilde{p} \geq p_0$ and $r = \tilde{r} \geq r_0$ with at least one of the inequalities strict. These are all models that are larger than the true model and nest the true model. The probability limit of $\ln|\frac{1}{T}\hat{U}'_{\tilde{p},\tilde{r}}\hat{U}_{\tilde{p},\tilde{r}}|$ for these models is the same as the probability limit of $\ln|\frac{1}{T}\hat{U}'_{p_0,r_0}\hat{U}_{p_0,r_0}|$. However, we know that $T(\ln|\frac{1}{T}\hat{U}'_{p_0,r_0}\hat{U}_{p_0,r_0}| - \ln|\frac{1}{T}\hat{U}'_{\tilde{p},\tilde{r}}\hat{U}_{\tilde{p},\tilde{r}}|)$ is the likelihood ratio statistic of testing general linear restrictions that reduce the $\tilde{p},\tilde{r}$ model to the p_0, r_0 model. Since these restrictions are true, $T(\ln|\frac{1}{T}\hat{U}'_{p_0,r_0}\hat{U}_{p_0,r_0}| - \ln|\frac{1}{T}\hat{U}'_{\tilde{p},\tilde{r}}\hat{U}_{\tilde{p},\tilde{r}}|) = O_p(1)$. While the reward for better fit from larger models is bounded in probability, the penalty terms for extra parameters increases without bound. Hence, the probability of choosing a larger model that nests the true model goes to zero asymptotically. This completes the proof that the first step of the procedure consistently estimates p_0 and r_0.

For the consistency of the second step estimator of q_0, we note that Chao and Phillips (1999) show that the PIC can be written as the sum (Chao and Phillips, 1999, express PIC as product of the likelihood and penalty term, here we refer to the logarithmic transformation of the PIC expressed in their paper) of two parts, one that comprises the log-likelihood of q given p and its associated penalty, and the other that comprises the log-likelihood of p without any restrictions on q and a penalty term involving the lag-length. With similar steps one can write the PIC in our case as the sum of one part related to q given p and r and another that involves p and r. Hence, plugging in p and r that are estimated via another consistent procedure does not alter the consistency of the estimator of q. The main reason that the choice of p and r does not affect the consistency of q is that the smallest

$K - q_0$ sample squared canonical correlations between Δy_t and y_{t-1} converge to zero in probability and the remaining q_0 converge to positive limits, *regardless of any finite stationary elements that are partialed out.* Therefore, for a given (p, r) when $q < q_0$, T times the difference in log-likelihood values dominates the penalty term, and hence the probability of underpredicting q goes to zero and $T \to \infty$. Also, when $q > q_0$, T times the difference in log-likelihood values remains bounded in probability, but the magnitude of the penalty for lack of parsimony grows without bound as $T \to \infty$, therefore the probability of overestimating q goes to zero asymptotically also. Note that the fact that the asymptotic distribution of the likelihood ratio statistic is not χ^2 or that it may depend on nuisance parameters does not matter. What is important is that it is $O_p(1)$. Hence the second step produces a consistent estimator of q_0, and this completes the proof.

Remark 3 *The above proof is not exclusive to HQ and applies to any model selection criterion in which $c_T \to \infty$ and $\frac{c_T}{T} \to 0$ as $T \to \infty$, where c_T is the penalty for each additional parameter in the first stage of the procedure. The consistency of model selection criteria with this property for determining p in vector autoregressions has been established in Quinn (1980), and in autoregressions with unit roots in Paulsen (1984) and Tsay (1984). Consistency of such criteria for selection of cointegrating rank q and the lag order p has been established in Gonzalo and Pitarakis (1995) and Aznar and Salvador (2002). Consistency of PIC for selection of cointegrating rank q and the lag order p has been established in Chao and Phillips (1999). The contribution here is proving the consistency when r is added to the set of parameters to be estimated, and showing that this can be achieved with a two-step procedure.*

Remark 4 *As with all models selected with any consistent model selection criterion, the warning of Leeb and Potscher (2005) applies to models selected with our procedure as well in the sense that there is no guarantee that any inference made based on asymptotic distributions conditional on p, q, r selected by this procedure will necessarily be more accurate than that based on an unrestricted autoregression of order $p_{\max}$.*

Remark 5 *Let $\tilde{\alpha}_1$ be a full rank $K \times (K - r_0)$ matrix such that $\tilde{\alpha}_1'[\ \Gamma_1 \quad \Gamma_2 \quad ... \quad \Gamma_{p_0}\ ] = 0$. Such a matrix exists because rank $[\ \Gamma_1 \quad \Gamma_2 \quad ... \quad \Gamma_{p_0}\ ] = r_0$ but it is not unique. We can augment $\tilde{\alpha}_1$ with r_0 additional linearly independent vectors arranged as columns of matrix $\tilde{\alpha}_2$ to form a basis for $\mathbb{R}^n$, and to achieve uniqueness we can choose these matrices such that $(\tilde{\alpha}_1 \vdots \tilde{\alpha}_2)'\Omega(\tilde{\alpha}_1 \vdots \tilde{\alpha}_2) = I_K$. The DGP can be alternatively written as*

$$\begin{aligned}
\tilde{\alpha}_1'\Delta y_t &= c_1 + \Pi_{(1)} y_{t-1} + \eta_{(1),t} \\
\tilde{\alpha}_2'\Delta y_t &= c_2 + \Pi_{(2)} y_{t-1} + \Gamma_{(2),1}\Delta y_{t-1} + \Gamma_{(2),2}\Delta y_{t-2} + \cdots + \Gamma_{(2),p_0}\Delta y_{t-p_0} + \eta_{(2),t}
\end{aligned}$$

where for any vector or matrix X, $X_{(i)} = \tilde{\alpha}_i' X$, $i = 1, 2$. *While we have presented the model selection criteria as penalised log-likelihoods and have referred to maximum likelihood estimators and likelihood ratio tests in our proof to conform with the previous literature, all arguments could be phrased in the context of GMM estimation of the above structural model and test statistics for testing overidentifying restrictions in the first block of this structure (Anderson and Vahid, 1998). Therefore, there is no need for any assumption of normality at any stage.*

C Tables

Table 1: Performance of IC(p, r, q) in a $(1, 1, 2)$ design and its comparison with the usual application of the Johansen method

	T=100				T=200				T=400			
	$q < q_0$	$q = q_0$	$q > q_0$	Total	$q < q_0$	$q = q_0$	$q > q_0$	Total	$q < q_0$	$q = q_0$	$q > q_0$	Total
AIC												
$r < r_0$	0,0,0	2,0,0	4,0,0	6,0,0	0,0,0	1,0,0	1,0,0	2,0,0	0,0,0	0,0,0	0,0,0	1,0,0
$r = r_0$	0,0,0	0,22,9	0,31,13	0,54,23	0,0,0	0,26,7	0,37,10	0,63,17	0,0,0	0,29,6	0,38,10	0,67,15
$r > r_0$	0,0,0	0,5,3	0,7,4	0,11,6	0,0,0	0,6,2	0,7,3	0,13,5	0,0,0	0,5,2	0,8,2	0,13,4
Total	0,0,0	2,27,12	4,38,17	6,65,29	0,0,0	1,32,9	1,44,13	2,76,22	0,0,0	0,34,8	0,46,12	1,80,19
AIC+J	1,9,1	10,41,3	6,26,3	17,76,7	0,2,0	4,53,3	2,34,2	6,89,5	0,0,0	1,55,3	1,38,2	2,94,4
HQ												
$r < r_0$	0,0,0	10,0,0	8,0,0	19,0,0	0,0,0	5,0,0	3,0,0	8,0,0	0,0,0	2,0,0	1,0,0	3,0,0
$r = r_0$	0,3,0	0,39,3	0,30,3	0,72,6	0,1,0	0,52,2	0,33,1	0,86,3	0,0,0	0,62,1	0,31,1	0,94,2
$r > r_0$	0,0,0	0,2,0	0,1,0	0,3,0	0,0,0	0,2,0	0,1,0	0,2,0	0,0,0	0,1,0	0,1,0	0,2,0
Total	0,3,0	10,41,4	8,31,3	19,75,6	0,1,0	5,54,2	3,34,1	8,89,3	0,0,0	2,63,1	1,32,1	3,95,2
HQ+J	2,8,0	20,34,0	14,22,0	37,63,0	0,2,0	11,48,0	8,31,0	19,81,0	0,0,0	4,55,0	3,38,0	7,93,0
SC												
$r < r_0$	3,0,0	24,0,0	9,0,0	36,0,0	0,0,0	15,0,0	4,0,0	19,0,0	0,0,0	7,0,0	1,0,0	8,0,0
$r = r_0$	0,8,0	0,42,1	0,13,0	0,63,1	0,4,0	0,63,0	0,14,0	0,80,0	0,1,0	0,79,0	0,12,0	0,92,0
$r > r_0$	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0
Total	3,8,0	24,42,1	9,13,0	36,63,1	0,4,0	15,63,0	4,14,0	19,81,0	0,1,0	7,79,0	1,12,0	8,92,0
SC+J	4,5,0	31,23,0	24,14,0	58,42,0	0,2,0	22,36,0	16,23,0	38,62,0	0,0,0	11,47,0	9,33,0	20,80,0
PIC												
$r < r_0$	7,0,0	24,0,0	1,0,0	32,0,0	1,0,0	14,0,0	0,0,0	15,0,0	0,0,0	5,0,0	0,0,0	5,0,0
$r = r_0$	0,14,0	0,52,0	0,2,0	0,68,0	0,6,0	0,77,0	0,2,0	0,85,0	0,2,0	0,92,0	0,1,0	0,95,0
$r > r_0$	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0
Total	7,14,0	24,52,0	1,2,0	32,68,0	1,6,0	14,77,0	0,2,0	15,85,0	0,2,0	5,92,0	0,1,0	5,95,0
HQ-PIC												
$r < r_0$	4,0,0	14,0,0	1,0,0	19,0,0	0,0,0	8,0,0	0,0,0	8,0,0	0,0,0	3,0,0	0,0,0	3,0,0
$r = r_0$	0,15,1	0,54,5	0,2,0	0,72,6	0,6,0	0,79,3	0,2,0	0,86,3	0,2,0	0,90,2	0,1,0	0,93,2
$r > r_0$	0,1,0	0,3,0	0,0,0	0,3,0	0,0,0	0,2,0	0,0,0	0,2,0	0,0,0	0,2,0	0,0,0	0,2,0
Total	4,15,1	14,57,5	1,2,0	19,75,6	0,6,0	8,81,3	0,2,0	8,89,3	0,2,0	3,92,2	0,1,0	3,95,2

Note: The total of the three entries a, b, c in each cell show the percentage of times the selected model falls in the category identified by the column and row labels. Entry a shows the percentage where $p < p_0$, b shows the percentage where $p = p_0$ and c the percentage where $p > p_0$. The row labeled X+J shows this information for the method commonly used in practice, where the lag-length p is chosen by model selection criterion X, and then the Johansen procedure is used for determining q.

Table 2: Performance of IC(p, r, q) in a $(2, 1, 1)$ design and its comparison with the usual application of the Johansen method

	T=100				T=200				T=400			
	$q < q_0$	$q = q_0$	$q > q_0$	Total	$q < q_0$	$q = q_0$	$q > q_0$	Total	$q < q_0$	$q = q_0$	$q > q_0$	Total
AIC												
$r < r_0$	1,0,0	1,0,0	1,0,0	3,0,0	0,0,0	1,0,0	1,0,0	2,0,0	0,0,0	0,0,0	1,0,0	1,0,0
$r = r_0$	0,0,1	1,11,4	4,34,14	6,45,19	0,0,1	1,14,4	2,41,11	3,57,15	0,0,1	0,16,3	1,44,10	1,60,14
$r > r_0$	1,1,3	1,3,2	2,9,6	3,13,11	0,0,3	1,3,3	1,8,5	2,11,10	0,1,2	1,3,3	1,8,5	2,11,11
Total	2,1,4	3,14,6	7,43,20	12,58,30	0,0,4	3,17,7	4,49,16	7,68,25	0,1,3	1,19,6	3,52,15	4,71,25
AIC+J	2,8,1	23,43,4	6,11,2	32,62,6	0,1,0	11,68,4	4,2,13	13,82,5	0,0,0	4,74,4	1,15,1	4,91,5
HQ												
$r < r_0$	0,0,0	3,0,0	3,0,0	6,0,0	0,0,0	2,0,0	1,0,0	3,0,0	0,0,0	1,0,0	0,0,0	1,0,0
$r = r_0$	0,1,1	8,37,4	6,25,3	14,64,8	0,0,0	3,56,3	2,25,2	6,79,5	0,0,0	2,65,2	1,21,2	2,85,4
$r > r_0$	0,0,1	1,1,1	1,2,1	3,4,3	0,0,1	0,1,1	1,1,1	1,3,3	0,0,1	0,1,1	0,1,2	1,3,4
Total	0,1,2	12,39,5	10,27,4	23,67,10	0,1,1	6,57,3	4,27,3	10,82,8	0,0,1	3,66,3	1,22,4	4,88,8
HQ+J	3,5,0	47,25,0	14,6,0	64,36,0	0,1,0	32,51,0	7,9,0	39,61,0	0,0,0	12,71,0	3,15,0	15,85,0
SC												
$r < r_0$	2,0,0	10,0,0	3,0,0	15,0,0	0,0,0	7,0,0	1,0,0	8,0,0	0,0,0	3,0,0	0,0,0	4,0,0
$r = r_0$	2,8,0	21,42,1	3,6,0	26,55,2	0,3,0	12,71,1	1,4,0	13,77,1	0,0,0	8,86,0	0,2,1	4,90,1
$r > r_0$	0,0,0	0,0,0	0,0,0	1,0,0	0,0,0	0,0,0	0,0,0	1,0,0	0,0,0	0,0,0	0,0,0	0,1,0
Total	4,8,0	32,42,1	7,6,0	42,56,2	0,3,0	19,71,1	2,4,0	22,77,1	0,0,0	11,86,1	0,2,1	8,90,1
SC+J	5,2,0	62,7,0	22,2,0	89,11,0	0,0,0	55,26,0	14,5,0	69,31,0	0,0,0	34,48,0	8,10,0	41,59,0
PIC												
$r < r_0$	4,0,0	11,0,0	1,0,0	16,0,0	0,0,0	7,0,0	0,0,0	7,0,0	0,0,0	3,0,0	0,0,0	3,0,0
$r = r_0$	4,3,0	37,28,0	1,1,0	42,41,0	1,4,0	25,62,0	0,0,0	26,66,0	0,0,0	10,87,0	0,0,0	9,88,0
$r > r_0$	0,0,0	0,0,0	0,0,0	1,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0
Total	8,3,0	48,28,0	2,1,0	59,41,0	1,4,0	32,62,0	1,0,0	34,66,0	0,0,0	13,87,0	0,0,0	12,88,0
HQ-PIC												
$r < r_0$	1,0,0	5,0,0	0,0,0	6,0,0	0,0,0	3,0,0	0,0,0	3,0,0	0,0,0	1,0,0	0,0,0	1,0,0
$r = r_0$	2,13,3	12,49,4	0,1,0	14,63,7	0,4,1	5,77,4	0,0,0	6,79,5	0,0,0	2,86,4	0,0,0	2,86,4
$r > r_0$	1,2,2	2,2,1	0,0,0	2,4,3	0,0,1	1,2,2	0,0,0	1,3,3	0,0,0	1,2,4	0,0,0	1,2,4
Total	4,15,5	19,51,5	1,1,0	23,67,10	1,4,2	9,79,6	0,0,0	10,82,8	0,0,0	4,88,8	0,0,0	4,88,8

Note: The total of the three entries a, b, c in each cell show the percentage of times the selected model falls in the category identified by the column and row labels. Entry a shows the percentage where $p < p_0$, b shows the percentage where $p = p_0$ and c the percentage where $p > p_0$. The row labeled X+J shows this information for the method commonly used in practice, where the lag-length p is chosen by model selection criterion X, and then the Johansen procedure is used for determining q.

Table 3: Percentage improvement in forecast accuracy measures for possibly reduced rank models over unrestricted VARs in a (1,1,2) setting.

Horizon	T=100			T=200			T=400		
(h)	TMSFE	|MSFE|	GFESM	TMSFE	|MSFE|	GFESM	TMSFE	|MSFE|	GFESM
VECM(HQ-PIC) for all DGPs									
1	1.4 (44,46,10)[a]	3.8	3.8	1.4 (49,49,2)	4.0	4.0	0.9 (53,47,0)	2.7	2.7
4	0.7 (23,77,0)	1.6	3.7	0.7 (46,54,0)	2.4	10.2	0.3 (27,73,0)	1.1	6.3
8	0.7 (19,80,1)	1.8	-7.2	0.1 (5,91,4)	0.1	8.0	0.1 (4,96,0)	0.5	6.8
12	0.2 (3,93,4)	0.5	-19.4	0.4 (14,86,0)	0.9	7.8	0.1 (4,96,0)	0.2	6.6
16	0.2 (5,94,1)	0.6	-31.3	0.4 (18,82,0)	1.0	3.7	0.1 (4,95,1)	0.2	7.2
VECM(AIC+J) for all DGPs									
1	0.9 (28,63,9)	2.3	2.3	0.8 (30,67,3)	2.3	2.3	0.4 (27,71,2)	1.0	1.0
4	0.4 (14,86,0)	0.6	2.0	0.2 (13,86,1)	0.8	5.5	0.1 (8,92,0)	0.4	2.2
8	0.5 (21,78,1)	1.4	-5.5	0.0 (2,91,7)	-0.2	4.2	0.1 (2,98,0)	0.2	1.9
12	0.1 (5,92,3)	0.4	-12.5	0.2 (12,88,0)	0.5	4.1	0.0 (0,98,2)	-0.1	1.4
16	0.1 (5,92,3)	0.4	-20.4	0.3 (18,82,0)	0.7	1.5	0.0 (3,97,0)	0.0	1.8

VECM(HQ-PIC) are models selected by the model selection process proposed in Section 4.1 and estimated by the algorithm proposed in Section 3. VECM(AIC+J) are estimated by the usual Johansen procedure with AIC as the model selection criterion for the lag length.

[a] We perform Diebold and Mariano (1995) tests at the 5% level of significance for equal predictive accuracy between the reduced rank models and unrestricted VARs. For cell (x,y,z), y denotes the percentage of DGPs for which the Null of equal forecast accuracy is not rejected and entries x and z denote the percentage of DGPs for which the Null is rejected with a positive statistic (i.e., the reduced rank model is significantly more accurate than the unrestricted VAR) and a negative statistic (i.e., the reduced rank model is significantly less accurate than the unrestricted VAR) respectively.

Table 4: Percentage improvement in forecast accuracy measures for possibly reduced rank models over unrestricted VARs in a (2,1,1) setting.

Horizon	T=100			T=200			T=400		
(h)	TMSFE	|MSFE|	GFESM	TMSFE	|MSFE|	GFESM	TMSFE	|MSFE|	GFESM
VECM(HQ-PIC) for all DGPs									
1	7.8 (87,13,0)[a]	21.8	21.8	4.5 (90,10,0)	12.9	12.9	2.5 (95,5,0)	7.5	7.5
4	2.2 (69,31,0)	8.1	37.8	2.0 (78,22,0)	5.2	30.6	0.9 (47,53,0)	2.3	17.5
8	1.0 (24,76,0)	2.7	38.5	0.6 (22,78,0)	2.3	34.1	0.6 (32,68,0)	2.2	25.7
12	0.4 (12,87,1)	0.8	29.8	0.8 (27,73,0)	2.4	36.8	0.9 (82,18,0)	2.9	29.5
16	0.8 (16,84,0)	1.8	25.5	0.3 (16,59,25)	0.3	32.8	0.7 (39,61,0)	2.4	32.7
VECM(AIC+J) for all DGPs									
1	5.4 (81,19,0)	14.1	14.1	3.2 (81,19,0)	8.7	8.7	1.4 (72,28,0)	4.1	4.1
4	1.3 (29,71,0)	4.8	21.6	1.2 (61,39,0)	3.0	21.3	0.6 (35,65,0)	1.8	10.7
8	0.7 (15,85,0)	1.9	21.5	0.6 (23,77,0)	2.3	26.1	0.4 (14,86,0)	1.7	16.8
12	0.5 (11,89,0)	0.9	14.5	0.6 (19,81,0)	1.9	29.6	0.7 (65,35,0)	2.4	19.2
16	0.6 (13,87,0)	1.4	11.0	0.2 (16,84,0)	0.3	27.4	0.6 (38,62,0)	2.2	22.0

VECM(HQ-PIC) are models selected by the model selection process proposed in Section 4.1 and estimated by the algorithm proposed in Section 3. VECM(AIC+J) are estimated by the usual Johansen procedure with AIC as the model selection criterion for the lag length.
[a] Refer to note in Table 3.

Table 5: Percentage improvement in forecast accuracy measures for reduced ranked models and unrestricted VARs for Brazilian inflation.

Horizon	VECM(HQ-PIC) versus			VECM(AIC+J) versus			VECM(HQ-PIC) versus		
(h)	VAR in levels			VAR in levels			VECM(AIC+J)		
	TMSFE	|MSFE|	GFESM	TMSFE	|MSFE|	GFESM	TMSFE	|MSFE|	GFESM
1	36.9	69.6	69.6	18.1	22.8	22.8	18.8	46.8	46.8
	(**,**,**)			(-, *, -)			(*, *,**)		
4	32.4	45.2	91.0	11.8	28.4	-10.6	20.6	16.8	101.6
	(**,**,**)			(-,**, -)			(**,**,**)		
8	24.6	32.9	107.9	15.1	26.2	-11.8	9.5	6.7	119.6
	(*,**,**)			(-,**,**)			(**, *,**)		
12	33.6	38.4	120.3	25.6	29.7	1.0	8.0	8.7	119.3
	(*,**,**)			(*,**,**)			(-, -, -)		
16	36.4	40.2	142.7	29.0	34.5	39.2	7.4	5.7	103.5
	(-, *, *)			(**,**, *)			(-, -, -)		

VECM(HQ-PIC) is the model selected by the model selection process proposed in Section 4.1 and estimated by the algorithm proposed in Section 3. VECM(AIC+J) is the model estimated by the usual Johansen procedure with AIC as the model selection criterion for the lag length. See Section 7 for further details. The triplet $(\cdot, \cdot, \cdot)$ presents the results of tests for equal mean squared forecast errors predicting $\Delta \ln(\text{CPI-IBGE}_t)$, $\Delta \ln(\text{CPI-FGV}_t)$, and $\Delta \ln(\text{CPI-FIPE}_t)$ respectively. The symbols **, * and - denote, respectively, significance at the 5% level, at the 10% level, and not significant at the 10% level.

Table 6: Percentage improvement in forecast accuracy measures for reduced ranked models and unrestricted VARs for U.S. macroeconomic aggregates.

Horizon	VECM(HQ-PIC) versus			VECM(AIC+J) versus			VECM(HQ-PIC) versus		
(h)	VAR in levels			VAR in levels			VECM(AIC+J)		
	TMSFE	|MSFE|	GFESM	TMSFE	|MSFE|	GFESM	TMSFE	|MSFE|	GFESM
1	35.1	60.4	60.4	16.2	49.7	49.7	18.9	10.7	10.7
	(**,**,**)			(**,**, *)			(-, *,**)		
4	56.3	83.5	134.7	27.8	46.4	112.1	28.5	37.1	22.6
	(**,**,**)			(**,**, *)			(**,**,**)		
8	8.4	25.3	169.2	8.9	24.0	145.2	-0.5	1.3	24.0
	(**,**, -)			(**,**, -)			(-, -, -)		
12	1.5	20.0	176.3	2.6	21.8	172.1	-1.1	-1.8	4.2
	(*,**, -)			(*,**, -)			(-, -, -)		
16	3.6	26.0	147.3	4.5	27.1	160.1	-0.9	-1.1	-12.8
	(**,**, -)			(**,**, -)			(-, -, -)		

VECM(HQ-PIC) is the model selected by the model selection process proposed in Section 4.1 and estimated by the algorithm proposed in Section 3. VECM(AIC+J) is the model estimated by the usual Johansen procedure with AIC as the model selection criterion for the lag length. See Section 7 for further details. The triplet $(\cdot, \cdot, \cdot)$ presents the results of tests for equal mean squared forecast errors predicting $\Delta \ln (y_t)$, $\Delta \ln (c_t)$, and $\Delta \ln (i_t)$ respectively. The symbols **, * and - denote, respectively, significance at the 5% level, at the 10% level, and not significant at the 10% level.